Introduction and validation of a pairwise comparison scale for UX evaluations and benchmarking with preschoolers

Bieke Zaman
CUO/IBBT, KULeuven
Parkstraat 45 bus 3605
3000 Leuven, Belgium
+32 16 323651
bieke.zaman@soc.kuleuven.be

ABSTRACT
This paper describes the development and validation of a pairwise comparison scale for user experience (UX) evaluations with preschoolers. More particularly, the dimensionality, reliability and validity of the scale are discussed. The results of two experiments among almost 150 preschoolers show difficulties to measure user experience quantitatively as a multi-dimensional construct with preschoolers. In contrast, the results suggest that UX should be measured directly as a one-dimensional higher order construct when preschoolers are involved. The one-dimensional scale proposed in this paper, encompassing five overall UX items, proved to be internally consistent and valid providing evidence of a solid theory-based instrument to measure UX with preschoolers.

Keywords
UX benchmarking & evaluations, preschoolers, user experience, methods, pairwise comparison scaling

INTRODUCTION

Making decisions on the development and launch of new technologies has become very difficult. Nowadays, complex product characteristics matter as unique selling points to distinguish from competitors. Companies have to come up with innovative ideas that generate a noticeable user experience that totally fits today's users' unfulfilled, and often unspecified dreams and wishes. These user experiences are often hard to reveal, and even harder to measure. Nonetheless, the launch of a new or improved digital product totally depends on whether it leads to a better user experience than the previous versions or competitor's products. The importance of good benchmarking measurement tools for decision makers and designers (e.g. during prototyping evaluations) can no longer be neglected. Because of this need for benchmarking, closed-ended scaling measures are preferred over open-ended scaling measures to objectively conclude whether one product dominates over another. In the context of this paper, dominance will be analyzed in terms of preferences based on actual user experiences. Once this preference is known, more qualitative or open-ended scaling methods are needed to reveal the reasons of dominance.

CLOSED-ENDED SCALING

Many types of closed-ended scaling exist such as dichotomous (nominal) scales, multiple category scales, rating (ordinal) scales (e.g. Likert scale), semantic differential or visual analogue scales. From all these scales, *dichotomous or pairwise comparison scaling* seems the least cognitively demanding questioning style for young children. In a benchmarking context, it perfectly allows for revealing preferences across products. However, in order to prevent response biases [see e.g. 7] and provide rich input for product design and evaluation, we should think further than yes/no scales. That is why we started thinking of pairwise comparison scaling with response categories that correspond with the products being evaluated, and attitude question items related to preference judgments. On the one hand, asking for a preference between two objects is a less sensitive way to talk about unpleasant experiences with objects than holding a discussion on the unpleasant experience alone. Children might feel less guilty to declare that one technology caused negative user experiences when at the same time they can compensate this answer by revealing how the other technology did result in good user experiences. On the other hand, asking for a preference between two or more objects, anticipates for social desirable answers in which everything is judged as 'fun'. Typically, many HCI researchers mistakenly evaluate new technologies by letting children explore these technologies and reporting on the enthusiastic reactions [4]. If no proper evaluation methods or research set-up are used, researchers do not exactly know what the enthusiasm can be accounted for. Children in general enjoy the situation in which adults observe them playing, while receiving undivided attention and gifts for their play (i.e. the typical incentives that are given to test children to motivate their participation). Regarding *multiple category scaling*, the risk of involving too many objects to choose from exists which can make the technique too cognitively demanding for young children. Consequently, we took the decision not to involve more than two response categories and instead focus on the feasibility of pairwise comparisons first. Further, we also decided not to work with both *semantic differential scales* as well as *visual analogue scales* due to the likelihood of risks in response biases (e.g. caused by the cognitive complexity). Nevertheless, because of the lack of research

on semantic differential with preschoolers, we keep this technique in mind for further work. Finally, in contrast to nominal scales, *ordinal scales such as rating scales (or Likert scales)* have the advantage that they do not only allow measuring differences in user experiences ('which object is preferred?') but also the strength of differences ('how much was this object relatively preferred compared to the other?'). Although very pertinent for UX evaluations, allowing a variety of statistical tests, it is not possible to administer these ordinal scaling questions with preschoolers. Nevertheless, instead of totally neglecting ordinal scaling for evaluations with young children, we will discuss in further work how, apart from the questioning techniques, we might transfer the data analysis techniques of ordinal scaling to pairwise comparison scaling. In this paper, however, we will focus on the development of our pairwise comparison scale aimed at benchmarking preschoolers' user experiences with technologies.

DEVELOPMENT OF A PAIRWISE COMPARISON UX SCALE

In [8] we described how a literature review and qualitative data resulted in a preliminary five component classification of the construct of user experience: 1) challenge & control, 2) fantasy, 3) creative and constructive expressions, 4) social experiences, 5) body and senses. These categories have been repeatedly reported as explanatory factors for what children positively experience and thus like in technology. We hypothesize that these categories are correlated and together measure the underlying construct of user experience (UX). This classification should make from UX a construct that can be measured multi-dimensionally with children. In the following paragraphs, we test the multi-dimensionality, reliability and validity of the scale.

RESEARCH QUESTIONS

Reliability and UX as a multi-dimensional construct

A first aim of our study was to test the multi-dimensionality of the construct representing the five UX components that are found to be relevant for children. More particularly, we tested whether it is justified to assume that these criteria can be accounted for by one single higher-order construct 'UX'. Our first set of research questions is:

RQ1a: Can the multi-dimensionality of user experience, revealed through literature review and qualitative analysis, be validated through quantitative, exploratory principal component analysis?
RQ1b: What is the reliability (internal consistency) of the scale?

Validity of the UX scale

As a second aim of our study, we tested the construct validity of our pairwise comparison scale. One type of construct validity is the convergent validity, usually measured by investigating the relation of the scale to measures of similar, theoretically related constructs. Construct validity can be broken down into a second category as well: the criteria validity. Criteria validity is measured by investigating the relationships between the scale and an independent criterion that is previously found to be related to this construct. This makes us to formulate the second set of research questions related to the scale's validity:

RQ2a: How does our UX scale relate to measures of convergent validity?
RQ2b: How does our UX scale relate to measures of criterion validity?

METHOD

Test participants and test setting

Two experiments were set up to test our scale. In April 2007, 36 preschoolers (17 girls and 19 boys) participated in the first experiment. Their age varied between 46 and 80 months and with an average age of five years and a half (M: 65 months, SD: 8 months). The second experiment took place one year later. In August 2008, 113 preschoolers were involved in the second experiment with ages ranging from 33 to 90 months and with a mean age of 58,39 months or 5 years (SD: 14 months). There was an equal gender division (56 boys versus 57 girls). All children were in kindergarten and not yet literate. The child participants of the first experiment were tested in their natural environment, namely in the (old) kitchen of the kindergarten. A holiday play initiative was involved for the second experiment.

This-or-That experimental set-up

Our experiments were set up as a This-or-That within-subject experiment in which the user experience with two technologies was compared. The This-or-That method refers to a mixed-method approach consisting out of four phases in which children are invited individually to judge preferences on user experiences of two technologies.

1. Exploration phase: Children are given the chance to explore the technologies first. Before one can ask young children about their experiences with a technology, it should be ensured that these experiences are fresh in memory. Indeed, attitudes are expressed more easily if the information needed to formulate these, is salient [5]. Moreover, experience attributes are relatively abstract and can only be assessed after using the product [2]. In general, the more recent and important the attitude information, the more relevant the attitude becomes, and the easier one can formulate an attitude judgment. To rule out carryover effects, the order in which children experience the technologies is counterbalanced. Last but not least, this exploration phase is also playful opportunity for the child to get used to the research situation, including researcher, test setting and equipment, and gives a common subject for researcher and child to talk about.

2. Quantitative survey questionnaire: After both exploratory conditions are finished, a questionnaire is administered. This questionnaire is based on the pairwise comparison scaling technique and aimed at revealing which

technology dominates in terms of user experience. The method's name 'This-or-That' refers to the special interview technique the questionnaire brings along. This technique consists of asking direct questions to the preschooler, stimulating to make a choice between two conditions ('pairwise comparisons'). For discriminating between the two options, the researcher actively points to the two alternatives while prompting "This one or that one? That one or this one?" The child then indicates the preferred technology, simply by pointing. This special technique of pairwise comparisons in combination with the use of contextual data in an individual face-to-face interview and pointing situation is likely to reduce the typical cognitive and social issues that are often impeding research projects with young children.

3. Qualitative probing interview: In order to interpret, check and validate the results based on the questionnaire, a short qualitative interview is administered. During this interview, the researcher will probe into the reasons *why* the child chose one condition over one another according to the principles of the contextual laddering method [9], which is a specific type of an attribute elicitation technique.

4. *Free play option:* Optionally, the answers on the questionnaire can be checked against a free play choice at the end of the test. More particularly, the researcher can decide to let the child play one of the two technologies again, as a 'reward' for their good participation. This free play option is especially relevant to triangulate the data obtained through the This-or-That questionnaire and allows validating the results statistically. More particularly, the affective, evaluative judgments of preference are compared to the behavioural component of preference. If the child holds strong preferences for one technology, then (s)he is likely to choose and play that technology again. Strong attitudes are indeed found to be better predictors of behaviour than weak attitudes [5].

MEASURES

User Experience: The preliminary scale used in the first experiment, measured user experience via 20 items. More particularly, three out of five *specific* components of the UX construct were measured through 15 questions (five for each component). We did not include questions related to 'social experiences' and 'creative and constructive expressions' because they were not applicable to our test case. Further, five more *overall* UX questions were added (asking for the game that was 'most fun', the game that was most desirable to 'receive as birthday present', 'take home', 'play again' or that was 'a little bit stupid'), resulting in a questionnaire of four subscales (of which three specific subscales and one overall UX subscale), measuring 20 items. Because of the low internal consistency of the first specific subscales (see further), we shortened the UX scale to the five more overall UX items in the second experiment. The UX scale of the second experiment thus consisted of the same items as the overall subscale of the first experiment and also generated a high internal consistency (see further).

Behavioural preferences: these were measured by one (experiment 2) or two (experiment 1) free play options, either immediately after the test or about two weeks later. In these situations, children were encouraged to select only one of the two conditions to play again as a 'reward for good participation' ('free play option').

Usability: In order to test our scale's criterion validity, we measured the relationship between 'usability' and user experience in our first experiment. Relying on ISO's 1997 definition of usability, we measured efficiency by the time necessary to complete the game. Effectiveness was measured by the number of subtasks successfully completed in the game and whether the child succeeded in finishing the game.

Qualitative data: Besides quantitative measurements, qualitative material was gathered as well. We video-recorded interaction styles and comments uttered by the preschooler when playing the games. Only after the complete test was finalized (i.e. playing the two conditions and answering the UX questions) the facilitator would follow up on this qualitative information and ask the preschooler to explain a little more on exactly *why one condition was chosen over one another* according to the contextual laddering method [9]. The qualitative data was important to test the accuracy of children's responses and resulted in a more rich understanding of the user experiences.

RESULTS

RQ1A: Multi-dimensionality of the construct: The first aim of our research (RQ1a) was to check whether the multi-dimensionality of UX revealed through literature review and qualitative analysis, could be validated through quantitative research. The results of the principal component analysis, however, suggest that it is not possible to measure the multi-dimensionality of the UX construct with young children (most probably due to their cognitive limitations). More particularly, the results of our principal component analysis could *not* confirm that for each component, the corresponding variables would correlate, resulting in three clear factors related to 'fantasy', 'body & senses' and 'challenge and control'. This makes us wonder whether UX should then be measured on a higher level – one-dimensionally-, directly referring to the overall user experience. Reliability analysis, as described in the next paragraph, helped us to decide on this issue.

RQ1B: Reliability of the scale - internal consistency: Principal component analysis of the results of our first experiment made us drop the three specific subscales in favour of selecting only the overall UX subscale. The questionnaire of 20 items was thus reduced to the following five items, measuring user experience on a higher level as a one-dimensional construct: 1. "Show me which product was most fun?" 2." Show me which product would you

like to receive as a gift?" 3. "Show me which product would you like to take home?"4. "Show me which product would you like to play again?"5. "Show me which product you found a little bit stupid?" In order to have an internal consistent scale encompassing these five items, we had to test the reliability. The reliability analysis of the overall UX scale in our first experiment, lead to a high Chronbach's alpha of .882 (M=7.91, SD=2.050). The same scale had an overall reliability score of Chronbach's alpha .797 in the second experiment (M=6.147, SD=1.569).

In sum, three arguments were put forth to select only the overall UX scale and drop the other subscales relating to the specific UX components of our preliminary UX classification. Firstly, exploratory principal component analysis did not reveal the expected factors. Secondly, the Chronbach's alpha of the subscales relating to the specific UX components, were too low. Finally, a last argument to select only the overall UX scale, consisting of five items, comes from developmental psychology: the fewer the number of questions, the more the questionnaire would be adapted to the cognitive and motivational capabilities of preschoolers. However, we had to check whether our limited scale would still be valid, which is discussed in the next paragraphs.

RQ2A: Convergent validity of the scale: In our experiments, the convergent validity was assessed by comparing scores on the UX scale to the free play option. The results of our experiments showed a significant correlation between the first free play moment and the scale (r=.570, N=33 for the first experiment, and r=5.81, N=111 for the second experiment, both Kendall's tau at the p<.01 level). As for the first experiment, there was also a correlation between the scale and the second free play moment with a Kendall's tau of r=.541 (p<.01, N=29).

RQ2B: Criterion validity of the scale: As for the scale's criterion validity, many research papers report on the relationship between usability and user experience [3,6]. The criterion validity was tested in our first experiment. The results are in line with previous studies. We indeed found significant correlations between the results on our UX scale and the usability of the game. More details on these correlations are discussed in [1]. In sum, the correlations between the UX scale and usability in our first experiment suggest a good criterion validity of our scale.

DISCUSSION

In this paper, we briefly explained how we developed and validated a pairwise comparison scale to measure and benchmark young children's user experience (UX) with digital technologies. The dimensionality, reliability and validity of the scale's items and components were tested in two experiments among almost 150 preschoolers. The results of our experiments reveal a one-dimensional scale encompassing five overall UX items that proved to be internally consistent and valid, thus promoting pairwise comparison scaling as a solution to perform quantitative UX evaluations with preschoolers.

Since this paper is only a concise introduction to UX measurements with preschoolers, more papers will definitely follow that go more deeply into the topics that arose in the context of this paper. For instance, follow-up papers will a) discuss the This-or-That method and its scale or interview techniques in more detail, b) elaborate on the appropriateness of different closed-ended scales for preschoolers, c) discuss the feasibility of experiments in which more than two alternatives are compared, d) explain how the advantages of dichotomous scaling questioning can be combined with the principles of ordinal scaling data analysis and e) describe the contextual laddering method with children more in detail.

REFERENCES

1. Abeele, V., Zaman, B., and Vanden Abeele, M. The Unlikeability of a Cuddle Toy Interface. *Book Chapter in Lecture Notes in Computer Science Volume 5294/2008*, Berlin: Springer: (October 2008), 118-131.
2. Bech-Larsen, T., and Nielsen, N. A comparison of five elicitation techniques for elicitation of attributes of low involvement products. *Journal of Economic Psychology, 20,* (1999), 315-341.
3. Hassenzahl, M., and Tractinsky, N. User experience – a research agenda. *B&IT, 25,2* (2006) UK: Taylor & Francis, 91-97.
4. Read, J. Validating the Fun Toolkit: an instrument for measuring children's opinions of technology. *Cognition, Technology & Work 10,2* (2008, April), Springer, 119-128.
5. Sudman, S., and Bradburn, N. *Response effects in surveys. A review and synthesis.* Aldine, Chicago, 1974.
6. Tractinsky, N. Aesthetics and apparent usability: Empirically assessing cultural and methodological issues. *Proceedings of CHI* (1997), 115-122.
7. Youngman, MB Designing questionnaires. In: Bell, J., Bush, T., Fox, A., Goodey, J., and Goulding, S. (eds) *Conducting small scale investigations in education management*. Harper and Row, London, 1984, 156–176.
8. Zaman, B., and Abeele, V. Towards a Likeability Framework that meets Child-Computer Interaction & Communication Sciences. *Proceedings of IDC* (2007), 1-8.
9. Zaman, B. Introducing contextual laddering to evaluate the likeability of games with children. *Cognition, Technology & Work 10,*2 (2008, April), Springer, 107-117.